Structuring the Blogosphere on News from Traditional Media

Georgios Petasis

Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research (N.C.S.R.) "Demokritos"
GR-153 10, P.O. BOX 60228, Aghia Paraskevi, Athens, Greece
petasis@iit.demokritos.gr

Abstract. News and social media are emerging as a dominant source of information for numerous applications. However, their vast unstructured content present challenges to efficient extraction of such information. In this paper, we present the SYNC3 system that aims to intelligently structure content from both traditional news media and the blogosphere. To achieve this goal, SYNC3 incorporates innovative algorithms that first model news media content statistically, based on fine clustering of articles into so-called "news events". Such models are then adapted and applied to the blogosphere domain, allowing its content to map to the traditional news domain. In this paper an unsupervised approach to do-main adaptation is presented, which exploits external knowledge sources in order to port a classification model into a new thematic domain. Our approach extracts a new feature set from documents of the target domain, and tries to align the new features to the original ones, by exploiting text relatedness from external knowledge sources, such as WordNet. The approach has been evaluated on the task of document classification, involving the classification of newsgroup postings into 20 news groups.

1 Introduction

News content in the internet, available through both traditional news media portals and the blogosphere, constitutes valuable information to both professionals and casual internet users, who however can be inundated by its vast amount. Clearly, such information could be much more useful if presented and delivered in a well-structured way. Many attempts, taking the form of either research projects or commercial solutions, have been made to provide centralised repositories of such content [1–3]. However, to date, there exists no integrated system that structures blog post content across these two broad sources of news information in parallel, capable to meet the requirements of a broad range of end users, such as professional journalists, communication experts, and citizen bloggers. The SYNC3 system [18] aims to fill this gap, efficiently structuring content from both domains, rendering it accessible, manageable, and re-usable.

The SYNC3 system is a solution for aggregating news from both traditional news media (i.e. news portals, etc.) and the blogosphere, providing the end users

Y.T. Demey and H. Panetto (Eds.): OTM 2013 Workshops, LNCS 8186, pp. 608–617, 2013.
© Springer-Verlag Berlin Heidelberg 2013

with sophisticated capabilities with respect to content structuring, management, and delivery. The methodology adopted applies the news domain structure derived from well-organised news portals to the less structured blogosphere. More specifically, SYNC3 automatically builds a news thematology, based on a statistical modelling approach that derives fine clusters of news articles, the so-called "news events". Subsequently, the system *adapts* the statistical news event models to the blogosphere domain, allowing the system to automatically find blog posts that comment on these events. Classifying blog posts into events extracted from news items can be easy, if the domain of both blog posts and news items are relatively similar. This can be the case for professional journalists who are also bloggers, as their writing style roughly remains the same when they write news items or blog posts. However, the vast majority of bloggers do not fall into this category, as they typically are individuals expressing personal thoughts, while their writing style may vary significantly from what is observed in the news. In order to associate blog posts from the latter category to the news, an adaptation of the classification model is required, in order to accommodate any possibly new writing styles. This process, known as domain adaptation, must extent a model for handling documents from a different domain (i.e., blog posts), without losing the ability to classify documents from the original domain (i.e., news items replicated in blog posts, or blog posts from journalists).

The portability of natural language processing (NLP) systems to new thematic domains is still a research area that attracts a significant research interest. During the last two decades, the use of machine learning has greatly improved the adaptability to new domains, or even languages. However, the vast majority of machine learning algorithms operate under a basic assumption: both the training and test data should use the same feature space, and follow the same distribution, suggesting that both should originate from the same thematic domain. When the distribution changes, the models must be re-generated from newly collected data. The adaptation can be separated into three large categories, according to the available data from the new domain. In supervised approaches, there is an adequate number of labelled data to train the model from scratch, on the new domain. When a limited number of labelled data are available, usually too few to train a model with satisfactory performance, along with unlabeled ones, the adaptation process is characterised as semi-supervised. Finally, unsupervised approaches must adapt their model to a new domain by learning solely from unlabelled examples.

Transfer learning or knowledge transfer is a research area, which tries to extract knowledge from previous experience and apply it on new learning tasks. Based on the idea that prior knowledge (i.e. identifying oranges) can be used on new tasks (i.e. identifying lemons), transfer learning researches three main central problems [22]: 1) how to extract the prior knowledge that is related, 2) how to represent the knowledge, and 3) how to apply the knowledge in the new learning task. Domain adaptation is a sub-category of transfer learning, where [17]:

- The source and target domains are different, but related.
- The source and target tasks are the same (i.e. classification or regression).
- Labelled examples are available for the source domain.
- Only unlabeled examples are available for the target domain.

In this paper, we propose a novel approach for the task of domain adaptation, in the context of document classification: we will try to classify (unlabelled) newsgroup posts from a target thematic domain D_T by performing model adaptation on a model acquired from labeled newsgroup posts belonging to a similar source domain D_S. Our method concentrates on the *feature space*, by trying to expand the features of the source domain with features that appear only in the target domain. Features that originate from the two different domains are aligned or linked to each other, through text relatedness. Text relatedness can take many forms, but we have opted for a simple relatedness measure, based on WordNet [15] synonymity. The rest of the paper is organized as follows: in section 2 related work is presented, where our method is compared to existing approaches. In section 3 our approach to model adaptation based on text relatedness is presented, while section 4 presents evaluation on the 20-newsgroup corpus [13]. Finally, section 5 concludes this paper and presents some future directions.

2 Related Work

The task of transfer learning can be defined as follows: given a source domain D_S, a source task T_S, a target domain $D_T \neq D_S$, and a target task T_T, transfer learning aims to learn a function f_T that accomplishes task T_T, by exploiting knowledge derived from D_S and T_S. A fairly recent overview of the area of transfer learning is given in the survey of [17], including the definition of transfer learning, its relation to traditional machine learning, a categorisation of transfer learning approaches, and practical applications of transfer learning. More recent approaches that target the task of domain adaptation can be found on the ACL 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP 2010) [8].

A lot of approaches exist that perform model adaptation in a fully supervised way (i.e. requiring labelled examples for both the source and target domains). For example, EASYADAPT [9] augments the source domain feature space using features extracted from labelled data in target domain. Prior work on semi-supervised approaches to domain adaptation also exists in literature. Recent work in domain adaptation has focused on approaches such as *self-training* and *structural correspondence learning* (SCL). The former approach involves adding self-labelled data from the target domain produced by a model trained in-domain [14]. The latter approach focuses on ways of generating shared source-target representations based on good pivot features [4,5,9]. However, the approach presented in this paper follows an *unsupervised approach*, thus requiring no labelled examples from the target domain. Unsupervised approaches try to exploit knowledge either from external knowledge sources, like our approach and [11], or from the distribution followed by the target domain [7,20]. The work presented in this paper can be

categorised as an "unsupervised feature construction" approach, according to [17]. Thus, approaches that try to extend a feature set through the unsupervised extraction of new features share some common ground with our approach. In [11] an approach that extracts new features by exploiting world knowledge is presented. World knowledge is represented through publically available ontologies, such as the Open Directory Project (ODP), where features from the source domain are mapped to appropriate ontology concepts, and "is-a" relations are exploited in order to acquire new features that augment the original feature set. Finally, the most appropriate features are selected through a feature selection phase. The work presented in [22] is also closely related to our approach: *feature correlation* is used in order to group features into *correlated groups*. For example, words like "orange", "lemon", "apple" and "pear" may often appear together in documents: aggregating them into a new correlated group "fruits", creates a new feature. If enough evidence exists in a document from the target domain (i.e. some of the features of the correlated group appear in the document), the feature that corresponds to the correlated group may help the task T_T in the target domain. In a sense, both approaches exploit information that can be characterised as "text relatedness" (or "feature relatedness"), as both "is-a" relations and correlation can be viewed as a relatedness measure between features.

However, our method has also some important differences with these two methods. Our text relatedness measure is based on synonymity, as provided by an electronic dictionary such as WordNet. An electronic dictionary may be an easier resource to find than an ontology or hierarchy, thus our approach may have a small advantage in initial requirements when compared to [12]. On the other hand, the calculation of feature correlation has no initial requirements in resources, but requires a corpus of adequate size, in order to extract the correlated groups. In addition, mining correlated groups may be computationally intensive if the feature set from the source domain is large enough (a problem tackled by limiting the source domain feature set to 2000 features, selected through mutual information, as reported in [22]). Finally, synonymity is a slightly more restricted text relatedness measure, compared to "is-a" relations (that can have many levels in the concept hierarchy) or correlation (which can relate possible unrelated features). Being a slightly more accurate text relatedness metric, it constitutes the need for feature selection, after the expansion of the source feature set, less important. In fact, our approach does not have a feature selection phase at all, in contrary to the two related approaches.

3 Domain Adaptation Based on Text Relatedness

The proposed methodology assumes a source domain D_S, a target domain $D_T \neq D_S$, a task T common for both domains, a feature space for the source domain $\mathcal{X}_S$, a label space $\mathcal{L}$ common for both domains, and a set of labelled examples originating from the source domain $L_S = \{X_1, \cdots, X_n\}$, where $X_i = \{x_1, \cdots, x_n,\ l_i\}$, $x_i \in \mathcal{X}_S$, $l_i \in \mathcal{L}$. In addition, our approach assumes a function $r\,(x_\alpha,\ x_\beta) \in \mathbb{R} : 0 \leq r \leq 1.0$, $x_\alpha, x_\beta \in \mathcal{X}_S,\ \mathcal{X}_T$, which decides if two features

are related, according to a text relatedness metric. Finally, a function $f_{\mathcal{X}_T}$ is assumed, that can extract a feature space $\mathcal{X}_T$ from the target domain D_T. The function $f_{\mathcal{X}_T}$ can be even a naive one, i.e. a function that returns all words in a corpus from the target domain D_T.

3.1 Text Relatedness Based on Synonymity

Our approach assumes a relatedness function $r\,(x_\alpha,\ x_\beta)$, that can compare two features (either from the source or from the target feature spaces), and return whether the two features are related or not. Although many relatedness metrics can be devised and used, we have opted for a simple one, based on synonymity. Assuming an electronic dictionary, which contains synonyms, text relatedness that is based on synonymity can be described with the following algorithm:

1. If x_α and x_β are the same, return 1.
2. Let S_α be the set of synonyms of x_α, and S_β the set of synonyms of x_β, according to the dictionary.
3. If $\mathrm{x}_\beta \in \mathrm{S}_\alpha$ or $\mathrm{x}_\alpha \in \mathrm{S}_\beta$, return 1.
4. If $\mathrm{S}_\alpha \cap \mathrm{S}_\beta \neq \emptyset$, return 1.
5. Else, return 0.

In simple words, our synonymity relatedness metric returns true, if the two features are synonyms, or when they have at least one common synonym. The electronic dictionary that has been chosen is WordNet [15], as has already been mentioned. It should be noted that all synonyms for all senses are treated equally, without performing any kind of word sense disambiguation [16], as is performed for example in the approach described in [12].

3.2 Extracting Features from the Target Domain

Our approach assumes that there is a function $f_{\mathcal{X}_T}$, which can extract features from the target domain D_T. Since no further requirements are assumed about this function, the function can be as naive or complex as the task T requires. We have considered a feature extraction procedure, which examines all documents of a corpus from the target domain D_T, and calculates the TF-IDF score for every word of the document. "Stop words" are rejected, and the rest of the remaining words are sorted according to their TF-IDF score, in a descending list. Then, an amount of the best scoring words, specified through a parameter θ (interpreted as a percent of the total words in a document), is extracted from each document, and added to the feature space that will be returned as the result.

3.3 Extracting New Features

Once we have a method for extracting possible new features from the target domain D_T, through the function $f_{\mathcal{X}_T}$, and a text relatedness metric $r\,(x_\alpha,\ x_\beta)$, we can apply these two functions in order to acquire a feature set from the target domain:

1. Let $\mathcal{X}_T^{Initial}$ be the feature space, as extracted from the target domain D_T by the function $f_{\mathcal{X}_T}$.
2. Each feature $x_s \in \mathcal{X}_S$ from the source feature set is compared to each feature $x_T \in \mathcal{X}_T^{Initial}$ in the extracted from the target domain feature set. The function $r(x_\alpha, x_\beta)$ is used for comparing the pair of features.
3. Features from the $\mathcal{X}_T^{Initial}$ that are not related to any feature in $\mathcal{X}_S$, are eliminated from $\mathcal{X}_T^{Initial}$, leading to a new feature space $\mathcal{X}_T^{Related}$.
4. As a final step, all features $x_T \in \mathcal{X}_T^{Related}$ are examined: every feature x_T that is related to more than one features in $\mathcal{X}_S$, is removed from $\mathcal{X}_T^{Related}$, leading to the final feature space that relates to the target domain $\mathcal{X}_T^{Final}$.

The result of this procedure, the final feature space that should be used for performing task T on the target domain D_T is the union of the two feature spaces: $\mathcal{X} = \mathcal{X}_S \cup \mathcal{X}_T^{Final}$.

3.4 Representing the Extracted Knowledge

The augmented feature space $\mathcal{X}$ that has been extracted as described in the previous subsection, contains all features of the source domain D_S, and new features from the target domain D_T, each of which is unambiguously related to a single feature from D_S. The only unsolved issue is how this augmented feature space is going to be represented as vectors, which can be used with a machine learning algorithm. Although this decision may rely on the particular machine learning algorithm that will be used, empirical evaluation suggested that the best alternative is to form "groups of features", where each old feature is replaced by a set of "related" features: the original one, plus the related ones from the target feature space, if they exist. This representation has been proved beneficial, at least for the task we have chosen to evaluate our approach (document classification), the chosen representation (bag-of-words) and the chosen classifier (kNN with $k = 1$ and cosine similarity as its distance metric).

4 Empirical Evaluation

In order to evaluate the algorithms proposed in the previous sections, we performed experiments on the 20-newsgroup dataset [13]: the 20-newsgroup dataset is a collection of approximately 20000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups, and is a standard evaluation corpus in many works related to domain adaptation or transfer learning. The task chosen for our empirical evaluation is document classification.

4.1 The 20-newsgroup Corpus

The 20-newsgroup corpus is preconfigured in training and testing material. Despite the fact that it is a popular evaluation corpus for domain adaptation approaches, it is unclear to us if all works that report results on the corpus use the same train/test partitioning, as different results are reported even for the base cases, as in [17] for example. In order to ease comparison with other approaches

Table 1. Corpus characteristics of the 20-newsgroup corpus

Pair	Posts for Domain	
	Source	Target
rec vs sci	4762	3169
rec vs talk	4341	2891
sci vs talk	4325	2880

we opted in using the predefined train/test segmentation of the corpus, as it is distributed. Regarding the task, we will limit evaluation to the three more popular evaluation pairs: "rec vs talk", "rec vs sci", "sci vs talk". The main idea behind the separation of these pairs, is that newsgroup posts from relevant but different newsgroups are put in the source/target domains. The "rec vs talk" class for example, may contain posts from the newsgroups "talk.politics.misc", "talk.politics.guns", "rec. motorcycles", and "rec.sport.hockey" as training material representing the source domain, while the test data (representing the target domain) may comprise from posts of the following newsgroups: "talk.politics.mideast", "talk.religion.misc", "rec.autos" and "rec.sport.baseball".

4.2 Experiment Setting

All posts in the three pairs of interest were pre-processed, in order for words to be recognised. A feature space from the posts constituting the training material was extracted, using the approach described in subsection 3.2, which extracts the top scoring words according to their TF-IDF weights, the number of which is controlled through a percentage of the total words of each post. This parameter was set to 0.003%, as it was found to roughly correspond to about one word from each post, leading for example to 4564 features for "rec vs sci", whose training material contains 4762 newsgroup posts. The reason behind this choice was to avoid possible over-fitting in the presence of too many features, and to provide our domain adaptation approach a chance to discover a large number of features from the target domain. As a measure of comparison, in [22] an initial feature space of 2000 features was selected.

Another point of interest is the choice of the machine learning algorithm, which will be used in order to learn a classification model. Support Vector Machines (SVMs) [6] are quite popular as a base case in model adaptation problems, since prior studies found SVMs to offer the best performance, at least for document classification using a bag-of-words representation [10, 21]. However, since our approach expands the feature space, we wanted to evaluate the effect of the augmented feature space with the least possible intervention from the chosen machine learning algorithm. Thus, we selected one of the simplest machine learning algorithms available, the k-nearest neighbour algorithm (kNN). kNN does not have a training phase, as it just classifies test instances using a similarity metric to measure distances from the training instances. In all experiments reported in this work, a kNN implementation was used with $k = 1$, and cosine similarity as the distance metric.

The bag-of-words representation was used for all experiments in this paper. Under this representation, each document (newsgroup post) is represented with a single vector, which has the same dimension as the feature space in use. The value for each feature is a real number, the TF-IDF weight of the feature in the document. The characteristics of the 20-newsgroup corpus, as well as evaluation results for the base classifier are shown in Tables 1 and 2 respectively.

4.3 Evaluation Results

The evaluation results of our approach are shown in Table 2. The upper part of Table 2 contains the evaluation results of our approach. The rows correspond to the examined pairs of newsgroups, while columns include information about the performance of the kNN classifier for the feature-space expansion phase. Evaluation results are presented in terms of precision, recall and F-measure (F_1). In table columns concerning recall, the improvement from the corresponding base case is also displayed, as difference between percentages. The lower part of Table 2 contains evaluation results from [19], where two model adaptation approaches were evaluated and compared with SVMs, used as a base case. As

Table 2. Evaluation results on domain adaptation for the 20-newsgroup corpus. Results from [19] are also shown for comparison purposes.

Feature space expansion based on text relatedness			
Pair	kNN ($k = 1$, cosine similarity)		
	Accuracy (base)	Accuracy (model adaptation)	
rec vs sci	40.07%	55.35% (+15.28)	
rec vs talk	51.78%	68.52% (+16.74)	
sci vs talk	41.67%	58.44% (+16.77)	
(Shi, Fan and Ren, 2008) [19]			
Pair	Accuracy (base/SVM)	Accuracy (TrAdaBoost)	Accuracy (AcTraK)
rec vs sci	59.1%	67.4% (+8.3)	70.6% (+11.5)
rec vs talk	60.2%	72.3% (+12.1)	75.4% (+15.2)
sci vs talk	57.6%	71.3% (+13.7)	75.1% (+17.5)

we can see from Table 2, the increase in performance achieved by our approach ranges from 15% (for "rec vs sci") to 19% (for "rec vs talk"). In comparison, the algorithm TrAdaBoost [7] achieved an increase ranging from 8% to 14%. The algorithm TrAdaBoost employs boosting in a semi-supervised approach, which exploits a small set of labelled data from the target domain, in addition to a large labelled data set from the source domain, in order to minimise the importance of labelled data from source domain (through weighting) whose distribution does not match the one of the target domain. AcTraK [19] achieves an additional improved of about 4% compared to TrAdaBoost, with the help of active learning in a semi-supervised approach, where labelled data may be asked when necessary. Our approach outperforms both approaches that represent the state of the art in the field, when applied on the task of document classification.

In addition, another interesting aspect of feature space expansion should be noted: the classifiers are able to provide an answer for a much larger number of newsgroup posts, even if the answer is not correct. For example, only 1571 (out of 3169) posts of the target domain contained features from the feature space of the source domain, in the case of the "rec vs sci" pair. After our approach has expanded the feature space with features from the target domain, 2309 posts of the target domain contained at least one feature from the augmented feature space, offering the possibility for classifying a larger number of posts.

5 Conclusions

In this paper, a domain adaptation approach was presented, that exploits text relatedness in the form of WordNet synonymity, in order to augment an initial feature space, derived from the source domain, with new features from the target domain. The proposed approached was empirically evaluated with the help of a manually annotated corpus. Evaluation results suggest that our approach can achieve an improvement comparable to other approaches that can be found in the bibliography, despite the fact that it employs kNN as its classifier to the task of document classification.

Acknowledgments. The author would like to acknowledge partial support of this work from the European Community Seventh Framework Programme, as part of the FP7 – 231854 SYNC3 project.

References

1. Europe Media Monitor (EMM) News Explorer, `http://emm.newsexplorer.eu`
2. Silobreaker Premium, `http://info.silobreaker.com`
3. Thoora Service, `http://thoora.com`
4. Ando, R.K.: Exploiting unannotated corpora for tagging and chunking. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, ACLdemo 2004. Association for Computational Linguistics, Stroudsburg (2004)
5. Blitzer, J., McDonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP 2006 (2006)
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995), `http://dx.doi.org/10.1007/BF00994018`
7. Dai, W., Yang, Q., Xue, G., Yu, Y.: Boosting for transfer learning. In: Proceedings of the 24th International Conference on Machine Learning, ICML 2007 (2007)
8. Daumé III, H., Deoskar, T., McClosky, D., Plank, B., Tiedemann, J. (eds.): Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing. Association for Computational Linguistics, Uppsala (2010)
9. Daumé III, H., Kumar, A., Saha, A.: Frustratingly easy semi-supervised domain adaptation. In: Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (2010)

10. Dumais, S., Platt, J., Heckerman, D., Sahami, M.: Inductive learning algorithms and representations for text categorization. In: Proceedings of the Seventh International Conference on Information and Knowledge Management, CIKM 1998, pp. 148–155. ACM, New York (1998), http://doi.acm.org/10.1145/288627.288651
11. Gabrilovich, E., Markovitch, S.: Feature generation for text categorization using world knowledge. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI 2005, pp. 1048–1053. Morgan Kaufmann Publishers Inc., San Francisco (2005)
12. Gabrilovich, E., Markovitch, S.: Feature generation for text categorization using world knowledge. In: IJCAI 2005, pp. 1048–1053 (2005)
13. Lang, K.: 12th International Conference on Machine Learning (ICML 1995) (1995)
14. McClosky, D., Charniak, E., Johnson, M.: Reranking and self-training for parser adaptation. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44, pp. 337–344. Association for Computational Linguistics, Stroudsburg (2006)
15. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM 38(11), 39–41 (1995)
16. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41(2), 10:1–10:69 (2009)
17. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)
18. Sarris, N., Potamianos, G., Renders, J.M., Grover, C., Karstens, E., Kallipolitis, L., Tountopoulos, V., Petasis, G., Krithara, A., Gallé, M., Jacquet, G., Alex, B., Tobin, R., Bounegru, L.: A System for Synergistically Structuring News Content from Traditional Media and the Blogosphere. In: Cunningham, P., Cunningham, M. (eds.) eChallenges e-2011 Conference Proceedings. IIMC International Information Management Corporation, Florence, Italy, October 26-28 (2011)
19. Shi, X., Fan, W., Ren, J.: Actively transfer domain knowledge. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 342–357. Springer, Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-87481-2_23
20. Thrun, S., Pratt, L.Y.: Learning to Learn. Kluwer Academic Publishers (1998)
21. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999, pp. 42–49. ACM, New York (1999), http://doi.acm.org/10.1145/312624.312647
22. Zhang, J., Shakya, S.S.: Knowledge transfer for feature generation in document classification. In: Proceedings of the 2009 International Conference on Machine Learning and Applications, ICMLA 2009, pp. 255–260. IEEE Computer Society, Washington, DC (2009)